Enhancing Scoring Reliability in Mid-Program Assessment of Design

Denny Davis, Michael Trevisan, Larry McKenzie
Washington State University
Steve Beyerlein
University of Idaho

Abstract

For the past six years, faculty across Washington State have worked to define and measure design competencies for the first two years of engineering and engineering technology degree programs. A three part performance-based assessment to assess student design capabilities at the mid-program level was developed for this purpose. This paper presents a pilot reliability study designed to enhance the consistency of scoring the three-part assessment. Toward this end, three raters participated in a multi-step procedure which included initial scoring of student work, reconciliation of differences among raters, revision of scoring criteria, and the development of decision rules to deal with student work difficult to score within the existing scoring criteria. Intraclass correlation coefficients were computed before and after this process, showing marked improvement of inter-rater reliability. The revised scoring criteria and decision rules offer potential for faculty to produce reliable scores for student design performance on constructed response items and tasks, a prerequisite to sound program decision making.

I. Introduction

The design capabilities among graduates from undergraduate engineering education programs continue to be a concern voiced by industry representatives. The need to improve design capabilities is further highlighted and motivated by requiring programs seeking accreditation through *ABET Engineering Criteria 2000*, to develop assessment competencies and a means to assess student design achievement[1]. In turn, this data is to be used as program feedback and when necessary, revisions to the program are to be made.

A prerequisite to effective use of assessments and sound programmatic decision making from assessment data is that achievement scores be obtained in a consistent manner. Consistent assessment data is referred to as “reliability” in the assessment literature, and signifies the extent to which the assessment is measuring achievement without error[2]. Despite the surge of interest in assessment processes within the engineering education literature in recent years, little discussion can be found regarding the quality of assessment data, such as reliability. The purpose of this paper is to illustrate one example for achieving consistent, reliable engineering design assessment results. The findings from this pilot study are preliminary. Multiple studies with all components of the assessment are now underway and may have ramifications for the nature and scope of the assessment. This paper provides a method for obtaining reliable data from a multi-faceted design assessment.

II. Transferable Integrated Design Engineering Education

For the past six years, a coalition of universities and community colleges in Washington state have led efforts of educators and industry representatives throughout the Pacific Northwest to establish effective processes for improving engineering design education. Known as the Transferable Integrated Design Engineering Education project or TIDEE, its central objective is to synthesize and institutionalize an outcomes-based engineering design education model across the northwest.

During the first three years of funding from the National Science Foundation, the TIDEE project developed design definitions and assessments for the first half of engineering programs, based on input from 2- and 4-year institutions across the nation. The genesis of this objective was twofold: (1) a concern for increased student design capabilities among industry stakeholders in the northwest, and (2) the realization that 26 community colleges in Washington state act as feeder schools to the 4-year institutions. The need to ensure comparable design education across the state became and remains the driving force behind this project[3,4].

The project defined three dimensions of the design learning domain that are fundamental to team-based engineering design: (a) design process, (b) teamwork, and (c) design communication. Students must master these three to be successful. Competencies to be obtained by students entering -- junior-level engineering programs (mid-program) are specified within each domain.

A three-component assessment is used to monitor student design capabilities at mid-program[5]. The first component is a set of short-answer constructed response (SCR) tasks that assess students' foundational knowledge about the design process, teamwork and design communication. Second, a performance assessment (PA) engages students in a team design activity. Students produce written documents that report team roles, the design process used, design requirements, and the design product. A reflective essay constitutes the third component of the mid-program assessment and provides further information on the team's design process and communication performance, and on member understanding of teamwork and communication processes. Separate scoring criteria depicting three levels of achievement accompany each task (See ref. 5 for a complete description of the assessments and associated scoring criteria.).

III. Establishing Reliability

Any decision made or informed by assessment data requires that scores be obtained in a consistent manner. Obtaining reliable, consistent scores obtained from subjective judgment, such as those used to score the constructed response tasks in the TIDEE mid-program design assessment, can be problematic and is a concern regardless of the assessment purpose or context. Specifically, differences in scoring between or among raters detract from consistency, and therefore, lower reliability estimates. However, work on rater scoring of performance assessments, essays, and constructed response tasks, has shown that high reliability can be obtained[6,7]. Requirements include clear scoring criteria, decision rules, and rater training.

There is not complete agreement in the literature regarding how best to estimate reliability in the context of subjective judgments of achievement. Some researchers use the Pearson Product

Moment (PPM) Correlation, computing the correlation of ratings between two raters, for example. However, differences between raters (variability) are disregarded using PPM correlations to estimate reliability. When differences in scoring exists among raters, the reliability estimate will be over estimated when using the PPM correlation to compute reliability.

A statistic, that does account for differences between or among raters and therefore has broad support for its use is the intraclass correlation coefficient[8,9]. The rationale for using the intraclass correlation coefficient (ICC) is that it systematically considers differences between raters when computing reliability. This statistic requires ANOVA methodology to compute and consequently, expresses reliability as deviations from the mean response, which is consonant with the historical definitions of reliability, particularly as defined in classical true score theory[10]. ICCs range in value from zero to 1.0 and can be interpreted as the percent of score variability due to the achievement domain the test or assessment is purported to measure; in this case, engineering design (See reference 8 for specific formulas).

Three decisions are central to the use of ICCs since the decisions dictate which of six types of ICCs will be computed. These questions are:

1. Will all raters rate all student work?
2. Are raters a random sample from a population of raters or are generalizations to a fixed set of raters?
3. Are reliability estimates and expectations based on one rater or the mean of more than one rater?

For this study, all raters rated all student work. Raters in this study were thought of as a random sample of raters from a population of raters. In addition, since this was a pilot reliability study, reliability estimates for scores for one rater as well as the mean of raters were computed.

IV. Scoring Process

In fall, 2000, the mid-program assessment was administered to a 139 mid-level students. This student sample was obtained from two institutions, constituted five classrooms, and represented mechanical engineering, electrical engineering, computer engineering, and biological systems engineering. The same individual administered the assessment to all students in a uniform manner.

Three raters were employed in this study. Two of the raters were engineering faculty members from different departments and institutions. The third rater was a research assistant with over 20 years industry experience managing engineering design projects.

A three-step scoring process was used to assess student work and test whether the scoring criteria could be applied reliably. This process was employed three times, one for each component of the mid-program assessment.

Step 1. A random sub-sample of student work ($n \approx 10$) was obtained from the original 139 students and scored collectively by the three raters. During this phase, raters openly discussed student work and whether or not the scoring criteria could be accurately applied. Differences in scores were reconciled and consensus scores obtained for the student sample. A key outcome from this step was a robust set of scoring criteria.

Step 2. A second random sub-sample of student work was obtained (n ≈ 30) and scored individually by each rater. Information obtained from the first step informed the independent scoring. ICCs were computed from the independent scores to estimate reliability of one rater and the reliability for the mean of three raters. After student work was scored individually and ICCs computed, raters again worked together to obtain consensus scores on student work through discussing student performance and reconciling score differences. In addition, decision rules were proposed to reduce inconsistencies in scoring student work difficult to accurately place within the existing scoring criteria. Some scoring criteria received minor modifications.

Step 3. The remaining sub-sample of student work (n ≈ 80) was scored independently by the three raters. ICCs were computed from the independent scores. Consensus scores were subsequently obtained for this sample of students, again by discussing student performance and reconciling score differences. Minor changes were also made to the scoring criteria and decision rules.

Figure 1 presents a detailed flowchart of the steps and sub-steps involved in the scoring process.

V. Findings

Table 1 shows the inter-rater reliability results before and after decision rules were developed and employed. The ICCs depict the reliability expected when generalizing to one rater and the mean of three raters, for each component of the assessment. The lowest reliability estimate for one rater before decision rules was 0.20 for the PA while the high was 0.62 for the essay. Low reliability estimates for the mean of three raters before decision rules was 0.43 for the PA. The high was 0.83 for the essay.

After decision rules, the lowest reliability estimate for one rater was 0.63 for the essay while the high was 0.77 for the SCR task assessing communication. The lowest reliability for the mean of three raters was 0.83 for the essay while the high was 0.91 for the SCR task assessing communication.

In general, the results show marked improvement in reliability for both ICCs, after the decision rules were developed and employed. The PA showed the greatest increase in reliability once decision rules were used.

VI. Discussion

Both ICCs across components showed acceptable magnitude for purposes of program decision making, and in some cases, exceed inter-rater reliability estimates commonly found with constructed response tasks[11]. The higher reliability estimates when generalizing to the mean of three raters is due to the more stable scoring of student achievement accrued when more raters are employed. However, this increase in reliability comes at a cost, since three raters (rather than one) are required to score all student work. Programs must consider and weigh the relative cost of one versus two or more raters and decide whether the higher reliability obtained with more raters is worth the investment of resources.

Advantages. In addition to the increase in reliability, the scoring process offers an instructional benefit. Specifically, the process of scoring student work, reconciling score differences among raters, revising score criteria, and developing decision rules, fosters a refined sense of engineering design among faculty participating in the scoring of student design work. Faculty are compelled to define design-related terms they've operated with in their classrooms, articulate theories, and defend concepts. Faculty come away from the experience with a better sense of engineering design, and as a consequence, are in a more informed position to teach these concepts. Thus, what initially started as a logistical requirement and duty by department faculty to score student work, is actually a powerful professional development exercise for participating faculty.

What also emerges from the scoring exercise is a collective sense of engineering design among participating faculty. Faculty now have a common vocabulary and set of concepts to discuss engineering design, its instruction, and student performance, across institutions. In turn, this common understanding has enhanced the comparability of design expectations and educational experiences across institutions, an outcome sought from the TIDEE project.

Disadvantages. The time commitment for initial steps of the process is substantial, although once the scoring criteria and decision rules are solidified and faculty have had practice with them, individual papers can be scored in less than three minutes each. Nevertheless, busy faculty may find it difficult to invest this much time given their other responsibilities. The task of scoring can be arduous as well. Faculty must be initially aware of the difficult work of scoring and be willing to stay the course for the process to work. In addition, getting faculty together periodically from geographically dispersed locations can be difficult and may be problematic for completing the task.

Additional Observations. Asking busy faculty to participate in an assessment development process such as the one described in this paper, justifies some form of compensation, since the time commitment requires more than the time allotted for a typical committee assignment. A reduction in course load seems a logical place to start and is a resource many departments can offer, even when budgets are limited. Moreover, those interested in expanding their scholarship can do so through documenting assessment development activities, proposing research questions tied to these processes and assembling research strategies to provide answers. The engineering faculty represented among the authors for this paper, provide a clear example. However, faculty with active research programs in other areas or faculty with other time consuming responsibilities must know that assessment development and maintenance activities will often conflict with these responsibilities, given the heavy time demands associated with these tasks.

VII. Conclusion

Defining engineering design and developing broad consensus on its features has been illusive for many in engineering education. Many models, theories, and frameworks currently exist, with no one model or theory leading in acceptance[12]. This fragmented view of engineering design complicates efforts to assess student design achievement consistently, threatening sound programmatic decisions. The process used in this pilot study suggests reliable data from design assessments is possible, despite the variety of engineering design models. By establishing a shared vision of design achievement, and formalizing this vision through scoring criteria and decision rules, engineering faculty can obtain high inter-rater reliability for constructed response design assessments. Engineering faculty can then be assured that assessment data used to inform decisions regarding the engineering design component of their programs are consistent, a prerequisite to sound programmatic decision making, now required by *ABET Engineering Criteria 2000*.

Figure 1.

Design Assessment Inter-rater Reliability Model

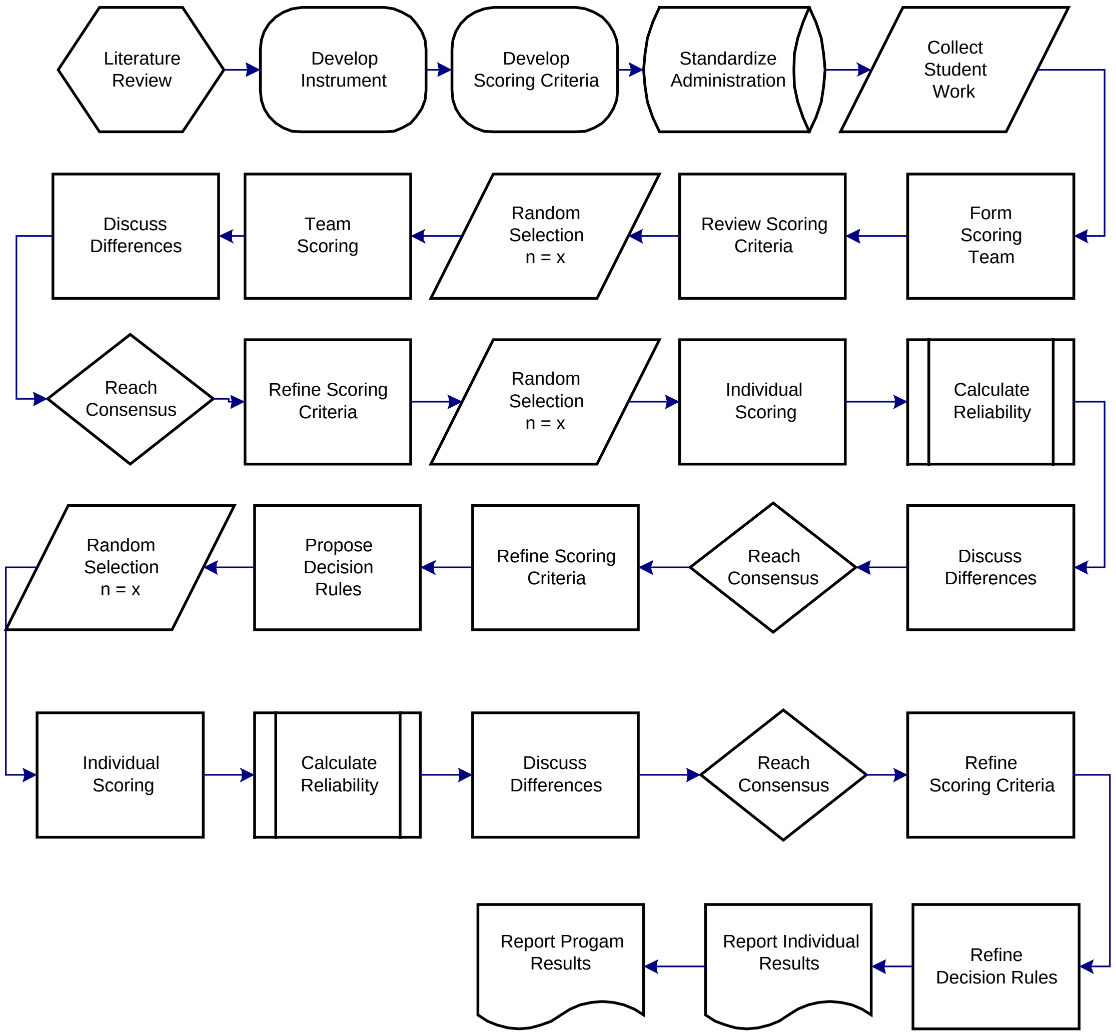

Table 1.

Intraclass Correlation Coefficients for Mid-Program Assessment (3 raters)

Stage in Scoring Process	n	Assessment Component	Inter-rater reliability estimate for one rater	Inter-rater reliability estimate for the mean of three raters
Before Decision Rules				
	28	SCR		
		Design	0.58	0.81
		Teamwork	0.56	0.80
		Comm	0.42	0.68
	35 (9 teams)	PA	0.20	0.43
	34	Essay	0.62	0.83
After Decision Rules				
	87	SCR		
		Design	0.69	0.87
		Teamwork	0.71	0.88
		Comm	0.77	0.91
	72 (16 teams)	PA	0.65	0.85
	83	Essay	0.63	0.83

Note: SCR = short-constructed response, Team = teamwork, Comm = communication, PA = Performance Assessment/Team Exercise; samples are random draws from student data set.

Bibliography

1. Engineering Accreditation Commission, *Engineering Criteria 2000, Criteria 2 and 3*, Accreditation Board for Engineering and Technology, Inc., Baltimore, MD, 2000.
2. Brennan, R. L. (2000). An essay on the history and future of reliability from the perspective of replications. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.
3. Davis, D.C., Gentili, K. L., Calkins, D. E., and Trevisan, M. S. (1998). Mid-Program assessment of team-based engineering design: Concepts, methods, and materials. Washington State University, Pullman, WA.
4. Davis, D.C., Gentili, K. L., Calkins, D. E., and Trevisan, M. S. (1998). Transferable integrated design engineering education – Final report. Washington State University, Pullman, WA.
5. Trevisan, M. S., Davis, D. C., Crain, R. W., Calkins, D. E., and Gentili, K. L. (1998). Developing and assessing statewide competencies for engineering design. *Journal of Engineering Education*, 87 (2), 185-193.
6. Herman, J. L., Aschbacher, P. R., and Winters, L. (1992). A practical guide to alternative assessment. Association for Supervision and Curriculum Development: Alexandria, VA.
7. Trevisan, M. S., Davis, D. C., Crain, R. W., Calkins, D. E., and Gentili, K. L. (1999). Designing sound scoring criteria for assessing student performance. *Journal of Engineering Education*, 88 (1) 79–85.
8. Armstrong, G. D. (1981). The intraclass correlation as a measure of interrater reliability of subjective judgments. *Nursing Research,* 30 (5), 314-315, 320A.
9. Tinsley, H. E. A., & Weiss, D. J., (1975). Interrater reliability and agreement of subjective judgment. *Journal of Counseling Psychology*, 22 (5), 358-376.
10. Trevisan, M. S. (1991). Reliability of performance assessments: Let's make sure we account for the errors. In P. Wolmut (Chair), *Measurement Issues In Performance Assessment*. Symposium conducted at the annual meeting of the National council on Measurement in Education, Chicago.
11. Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-303.
12. Wood, D. R. (2000). An evidence-based strategy for problem solving. *Journal of Engineering Education*, 89 (4) 443-459.

Author Biographies

MICHAEL S. TREVISAN is an associate professor in the Department of Educational Leadership and Counseling Psychology at Washington State University. Since 1997, he has served as the Director, Assessment and Evaluation Center within the department. He received a B.A., Mathematics from San Jose State University in 1983, M.Ed., Educational Psychology in 1988, and the Ph.D. degree in Educational Psychology in 1990, both from the University of Washington. His research interests include educational assessment, applied measurement and statistics, and program evaluation.

DENNY C. DAVIS is a professor in the Department of Biological Systems Engineering at Washington State University, where he has served as Department Chair since 2000. He served as Associate Dean, College of Engineering and Architecture, WSU, from 1986-1998. He has directed the TIDEE project since 1989. He received a B.S., Agricultural Engineering from Washington State University, and M.S. and Ph.D. degrees in Agricultural Engineering both from Cornell University.

LARRY J. MCKENZIE is a second year doctoral student in Educational Psychology at Washington State University. Prior to his graduate studies, he held various leadership and project management positions for 4 years in the U.S. Nuclear Navy, and 18 years with Duke Engineering & Services of Charlotte, NC. He received a B.A. in Chemistry from West Virginia University. His research interests include assessment in higher education and program evaluation.

STEVEN W. BEYERLEIN is an associate professor in the Department of Mechanical Engineering at the University of Idaho. He received a B.S.M.E. Mechanical Engineering from the University of Massachusetts in 1979, an M.S. Engineering, Dartmouth College in 1981and Ph.D. Mechanical Engineering, Washington State University in 1987. His research interests include development, testing, and modeling of catalytic ignition systems for spark-ignition and compression-ignition engines, alternate fuel usage and emissions testing, technology integration and performance monitoring for hybrid electric vehicles, application of educational research methods in engineering courses, and design and delivery of faculty development activities.